SAS® GLOBAL FORUM 2018

USERS PROGRAM

Ranking Between the Lines

A %MACRO for Interpolated Medians

By Joe Lorenz

April 8 – 11 | Denver, CO

#SASGF

Ranking Between the Lines: A %MACRO For Interpolated Medians

Joe Lorenz

Grand Valley State University

ABSTRACT

- This presentation goes through the process of creating a %Macro to find interpolated medians in BASE SAS. While medians are a great way to summarize the center of skewed data, but when collected data comes from an ordinal scale or is drawn from a very small range in possible values. Two medians can be exactly the same but come from data that is weighted very differently and because of that the median does not always accurately represent the center and shape of the data as accurately as it could. Interpolated Medians on the other hand can better represent whether the weight of the data is above or below the true median and therefore not only tell you something about the center of the median but also describe the shape of the data.
- This presentation will explain the origins, calculations and uses of interpolated medians. I will then show an example of where the interpolated succeeds where the regular median falls short by showing two sets of data with the same possible range of values but very different samples, and show that they have the same median and show how the interpolated median becomes much more descriptive and a better measure in this situation. I will then go through the code I used to develop the %MACRO to calculate interpolated medians, along with the tests I ran to validate it against known datasets and known interpolated medians to ensure that the %MACRO functions correctly and is an easy to use way to calculate interpolated medians in SAS®.

History & General Info

- Developed by J.P. Guilford in 1965 in his article *Fundamental Statistics in Psychology and Education*
- Interpolated Medians are used to help better reflect the center of distributions of data that come from a small range of values
- This will scale your median ± 0.5 of it's original value in order to show if the data is weighted majorly above and on the median or below
- This is extremely useful in survey data when you are working with ordinal values on a Likert scale
- Converts your ordinal median data to quantitative ratio data which can help give more accurate results when scoring rankings

When To Use

Interpolated medians are within ±0.5 of the value of the true median.

- If the value of the Interpolated Median is below the true median, it means the majority of the values are weighted below the true median.
- If the value of the Interpolated Median is above the true median, it means the majority of the values are above the true median.

Consider this example from the University of Michigan

Response	Question 1	Question 2
5 = Strongly agree	9	1
4 = Agree	10	10
3 = Neither agree nor disagree	0	6
2 = Disagree	1	1
1 = Strongly disagree	0	2

In the above Likert data the Median would both be the same for the two questions, with the median response being 4 – Agree.

- However for Question 1 the weight respondents clearly responded more in the Agree-Strongly Agree, so the interpolated median is 4.4.
- For Question 2 the data is weighed much more heavily in the Neither Agree-Degree – Agree, so the interpolated median 3.6.

If you wanted to know if raters were responding the two questions the same, using medians you would say yes believing the median is 4, however interpolated medians show a distinct difference.

Ranking Between the Lines: A %MACRO For Interpolated Medians

Joe Lorenz

Grand Valley State University

The Formula

M = Standard Median
IM = Interpolated Median
nl = Number of Responses < M
ng = Number of Response > M
ne = Number of Response = M

If ne is nonzero: $IM = M + \frac{ng - nl}{2ne}$

If ne is zero then: $IM = M$

The %MACRO

```
%macro intmed(dset, var);

***Finds the true median for the
Ordinal Data;
proc means data = &dset. median
noprint;
  var &var.;
  output out = outmed median=median;
run;
***Creates a macro variable for the
median;
data _null_;
  set outmed;
  call symput("Median", median);
run;
```

```
data transform;
  set &dset. end=last;

  ***Creates the variables to store
each variable above;
  retain  nl ne ng 0;
  median = &median.*1;

  ***Increments the count;
  if &var. < median then nl = nl+1;
  if &var. = median then ne = ne+1;
  if &var. > median then ng = ng+1;
```

The %MACRO

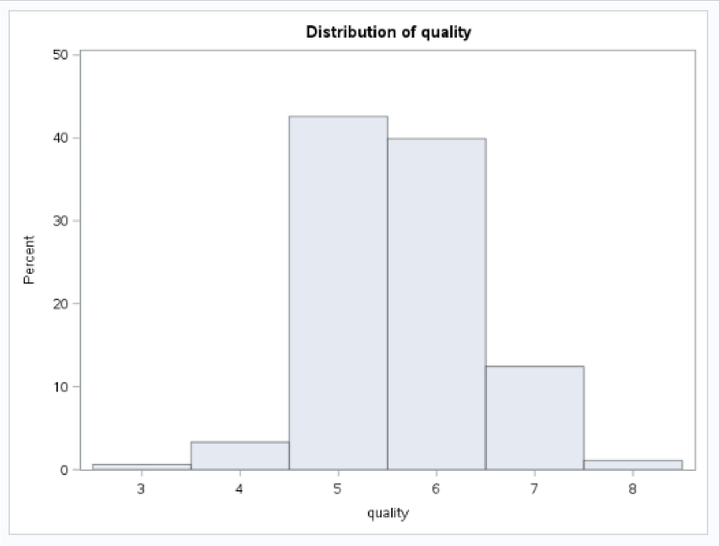
```
 ***Uses the formula to call the interpolated median;
  if last then do;
   if ne ne 0 then do;
     intmed = median + (ng-nl)/(2*ne);
   end;

   else if ne = 0 then do;
     intmed = median;
   end;
   output;
  end;

  label median = "Median"
        intmed = "Interpolated Median";
  keep median intmed;
run;

*Prints median and interpolated median;
proc print data = transform label noobs;
  title "Value for Median and Interpolated Median";
  var median intmed;
run;
title "";
%mend;
```

Ranking Between the Lines: A %MACRO For Interpolated Medians

Joe Lorenz

Grand Valley State University

Example



Value for Median and Interpolated Median

Median	Interpolated Median
6	5.58699

CONCLUSIONS

- Interpolated Medians were developed by Guilford in 1965 and has not been widely used since then, however through the use of a SAS %MACRO it is now easy to compute and obtain Interpolated Medians in SAS®
- Interpolated Medians more accurately reflect the area where the center of the data is while also taking into account where the data is weighted most heavily
- This method is more accurate when reporting medians for small ordinal scales, like are often found reported in Likert value based surveys.
- This macro can easily be used to either find global interpolated medians for an entire dataset or individual interpolated medians
- Great to be used with other survey analysis techniques since surveys employ Likert scales often
    - Interclass Correlation
    - Interrater Reliability

REFERENCES

P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.

Guilford, J. P. (1965). Measure of Central Value. Fundamentals of Statistics in Psychology, 4th(1), 49-55.

University of Michigan. (n.d.). Administration Evaluation. Retrieved February 24, 2018, from http://aec.umich.edu/median.php

SAS® GLOBAL FORUM 2018

April 8 – 11 | Denver, CO
Colorado Convention Center

#SASGF

Paper 2508-2018

Ranking Between the Lines: A %MACRO For Interpolated Medians

Joe Lorenz, Grand Valley State University

ABSTRACT

This paper goes through the process of creating a %Macro to find interpolated medians in BASE SAS®. While medians are a fantastic way to summarize the center of skewed data, when the collected data comes from an ordinal scale or is drawn from a very small range in possible values it becomes less useful. Two medians can be the same but come from data that is weighted very differently and because of that the median does not always accurately represent the center and shape of the data as well as it could. Interpolated Medians on the other hand can better represent whether the weight of the data is above or below the true median and therefore not only tell the researcher something about the center of the median but also describe the shape of the data.

This paper will explain the origins, calculations and uses of interpolated medians. It will also show an example of where the interpolated median is a better fit than the regular median on two sets of data, with the same median, with the same possible range of values but very different distribution and show how the interpolated median becomes much more descriptive and a better measure in this situation. This paper will then go through the code used to develop the %MACRO to calculate interpolated medians, along with an example demonstrating that the %MACRO functions correctly and is an easy to use method to calculate interpolated medians in SAS®.

INTRODUCTION

This paper will attempt to go through interpolated medians focusing on how to create a %MACRO to find an interpolated median for a variable using BASE SAS®. The focus of this paper will be on how and when to use interpolated medians, and interpreting the output from the %MACRO, as well as an example of using the %MACRO. For those who have not used interpolated medians before or are unsure of how interpolated medians are calculated there will be a brief introduction to interpolated medians. This introduction will cover the history of interpolated medians, a practical example of why interpolated medians have advantages over their un-interpolated counterpart, and the formula for interpolated median and how to calculate interpolated medians by hand.

AN INTRODUCTION TO INTERPOLATED MEDIANS

Interpolated medians were first introduced by Guilford in his 1965 paper *The Measures of Central Value*. Guilford discussed the interpolated median as a way of creating a more accurate median, by first finding the true median of the data, and then either increasing the value within a range ±0.5 (Guilford, 1965). This is useful if the data is not on a continuous scale, but instead has very low variability like ordinal data. Likert scale data and very small ordinal scales in particular can especially benefit from the extra level of sensitivity that interpolated medians have when describing the center of the data. This is done in order to account for if there are more values above or below the median, ignoring any observations in the data that fall exactly on the median. Using this methodology Guilford eventually developed the formula below to find the exact value of the interpolated median (1965).

M = Standard Median
IM = Interpolated Median
nl = Number of Responses < M
ng = Number of Response > M
ne = Number of Response = M

$$If\ ne\ is\ nonzero: \quad IM = M + \frac{ng - nl}{2ne}$$
$$If\ ne\ is\ zero\ then: \quad IM = M$$

Figure 1. The Formula for Interpolated Medians

To find the interpolated median the user must first find the true median of the data, this is represented by M in the formula above. After having the true median of the data, the user must compare all observations and identify whether they fall below (nl), on (ne), or above the true median (ng) (Guilford, 1965; University of Michigan, n.d.). If there are no values of the data that fall on the true median, then the interpolated median is the same value as the true median of the data. Otherwise the interpolated median is the median plus the number of responses above the median minus the number of responses below the median divided by 2 times the number of responses on the median.

Response	Question 1	Question 2
5 = Strongly agree	9	1
4 = Agree	10	10
3 = Neither agree nor disagree	0	6
2 = Disagree	1	1
1 = Strongly disagree	0	2

Figure 2. An Interpolated Median Example

To understand why the interpolated median is useful, consider the data above which reflects responses to two different questions both of which are recorded on the same 5-point Likert scale (University of Michigan, nd.). Both questions have a median response of 4, but as it is clearly seen on the figure above the distribution of these two variables are very different with Question 1 mainly having very favorable responses, with only one non-favorable response. Question 2 however has some favorable results, but also has a lot of very large amount of average responses, and even more non-favorable responses. Using the formula for interpolated medians however a difference can be seen that isn't present in the standard medians. The interpolated median for question 1 is 4.4, since this is above the median of 4 it means that the true value for the median is 4, but that there are more values above the median in the data than below. The interpolated median for question 2 is 3.6, which is below the true median, this means that the true median response is 4, but that there are more values below the median then above the median. These major differences between the true median and interpolated median show the benefit of using interpolated medians when the data comes from small ordinal scales. Now that we know how to calculate an interpolated median, we can develop a %MACRO to automate this process.

THE INTMED %MACRO

To create the %MACRO the user must first identify the dataset and the variable for which the user wishes to find the interpolated median of. The true median of the data is then found, using the MEANS procedure, outputted, and inserted into a macro variable, called "Median", in the null DATA step:

```
%macro intmed(dset, var);


***Finds the true median for the Ordinal Data;
proc means data = &dset. median noprint;
```

```
  var &var.;
  output out = outmed median=median;
run;

***Creates a macro variable for the median;
data _null_;
  set outmed;
  call symput("Median", median);
run;
```

The following DATA step then uses this %MACRO variable and a RETAIN statement to count how many observations for the user's variable fall above, below, and on the value of the median. Once it has finished processing through the data, this information is then used to calculate the interpolated median using the formula discussed in the previous section:

```
data transform;
  set &dset. end=last;

  ***Creates the variables to store each count of the levels of the ordinal
varaible;
  retain  nl ne ng 0;
  median = &median.*1;

  ***Increments the count;
  if &var. < median then nl = nl+1;
  if &var. = median then ne = ne+1;
  if &var. > median then ng = ng+1;

  ***Uses the formula to call the interpolated median;
  if last then do;
   if ne ne 0 then do;
     intmed = median + (ng-nl)/(2*ne);
   end;

   else if ne = 0 then do;
     intmed = median;
   end;
   output;
  end;

  label median = "Median"
```

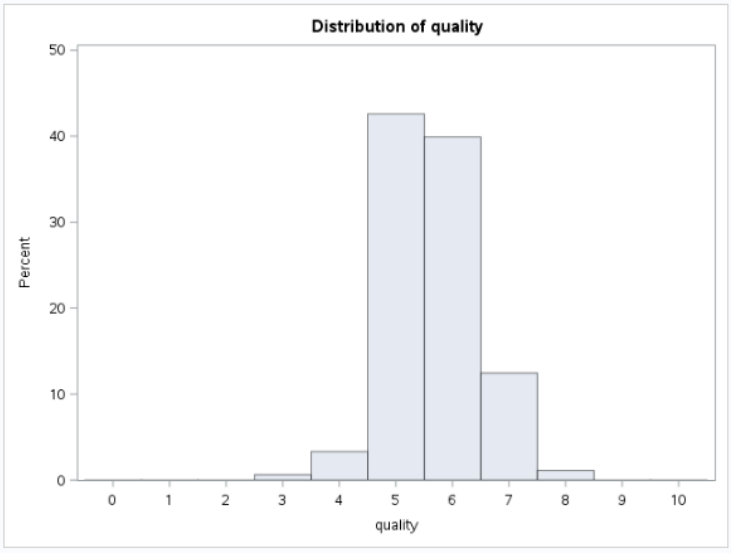
```
        intmed = "Interpolated Median";


  keep median intmed;
run;
```

The last thing that is done within this %MACRO is using the PRINT procedure to display the median and interpolated median side-by-side:

```
*Prints median and interpolated median;
proc print data = transform label noobs;
  title "Value for Median and Interpolated Median";
  var median intmed;
run;
title "";
%mend;
```

AN EXAMPLE

This data was made available thanks to Paulo Cortez, and the University of Minho in Portugal (Cortez, 2009). The data describes information on 1599 different red wines from the Minho region of Portugal, which accounts for about 15% of total wine produced in the country. The data includes many different continuous values for the different attributes of the wine including measure of acidity, sugar, chlorides, sulfur dioxide, density, pH, sulfates, and alcohol percentages. The variable being used in this example is the overall wine score, which is recorded on a 10-point ordinal Likert scale. The distribution and the original median of the red wine score was found using the UNIVARIATE procedure.



Basic Statistical Measures			
Location		Variability	
Mean	5.636023	Std Deviation	0.80757
Median	6.000000	Variance	0.65217
Mode	5.000000	Range	5.00000
		Interquartile Range	1.00000

Figure 3. Selected Wine Quality Output from PROC UNIVARIATE

Looking at the above distribution of the data it is clear to see that the true center of the data is somewhere in the 5 to 6 range since most of the data falls around here, however because there are a few raters rating around 7 or 8 they are pulling the true median from 5 to 6. Using the true median, the researcher would not know whether most of these values are weighted above or below the median. However, an

interpolated median would reflect that not only does the median for this data lie at 6, but also that most of the data is weighted below the true median.

Value for Median and Interpolated Median

Median	Interpolated Median
6	5.58699

Figure 4. Output from the %INTMED %MACRO

Looking at the above output from the INTMED %MACRO created above the macro confirms that the median value for the variable is the same at 6, but that the interpolated median is 5.58699. As talked about above since decimal value of the interpolated median is above 0.5, that means the true median is found by rounding up which would be 6, but that the weight of the data falls more below the median than above it. This means that while the median wine rating for these red wines is 6, more values are rated below 6 than above it.

CONCLUSION

This paper has demonstrated that interpolated medians can be a better representation of the true center of the data than the standard median. This is especially true when the origin of the data is ordinal in nature, like with a lot of survey data that records responses on small Likert scales. This will give the researcher a more accurate reading of the center of the data and can be especially helpful when one want to use the data to do more advanced analysis, like understand if raters are rating questions the same in inter-rater reliability analysis, seeing if questions are related using interclass correlation analysis. While interpolated medians used to involve calculating the value for the interpolated median by hand, using this %MACRO finding the interpolated median for a variable has never been easier.

REFERENCES

Guilford, J. P. (1965). Measure of Central Value. Fundamentals of Statistics in Psychology, 4th(1), 49-55.

P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis (2009). Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.

University of Michigan. (n.d.). Administration Evaluation. Retrieved February 24, 2018, from http://aec.umich.edu/median.php

ACKNOWLEDGMENTS

The author would like to acknowledge Grand Valley State University for his time at the Statistical Consulting Center where he first learned about interpolated medians. Specially Dr. Sango Otieno who suggested he create a %MACRO to perform and create interpolated medians, and Dr. Robert Downer, his faculty advisor at Grand Valley State University.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Joe Lorenz
lorenzj@mail.gvsu.edu
www.linkedin.com/in/joe-lorenz

```
***
The calls on the macro are the dataset that the variable is being stored in and the variable that is being analyzed;
%macro intmed(dset, var);

***Finds the true median for the Ordinal Data;
proc means data = &dset. median noprint;
  var &var.;
  output out = outmed median=median;
run;

***Creates a macro variable for the median;
data _null_;
  set outmed;
  call symput("Median", median);
run;

data transform;
  set &dset. end=last;

  ***Creates the variables to store each count of the levels of the ordinal varaible;
  retain  nl ne ng 0;
  median = &median.*1;

  ***Increments the count;
  if &var. < median then nl = nl+1;
  if &var. = median then ne = ne+1;
  if &var. > median then ng = ng+1;

  ***Uses the formula to call the interpolated median;
  if last then do;
   if ne ne 0 then do;
     intmed = median + (ng-nl)/(2*ne);
   end;

   else if ne = 0 then do;
     intmed = median;
   end;
   output;
  end;

  label median = "Median"
        intmed = "Interpolated Median";

  keep median intmed;
run;

*Prints median and interpolated median;
proc print data = transform label noobs;
  title "Value for Median and Interpolated Median";
  var median intmed;
run;
title "";
%mend;
```